From Terminology to Evaluation: Performance Assessment of Automatic Signature Verification Systems

Muhammad Imran Malik*† , Marcus Liwicki*
* *German Research Center for AI (DFKI GmbH)*
Knowledge Management Department, Kaiserslautern, Germany
{*firstname.lastname*}*@dfki.de*
† *Department of Computer Science, University of Kaiserslautern, Germany*

***Abstract*—This paper is an effort towards the development of a shared conceptualization regarding automatic signature verification systems. The requirements of both communities, Pattern Recognition and Forensic Handwriting Examiners, are explicitly focused. This is required because an increasing gap regarding evaluation of automatic verification systems is observed in the recent past.**

The paper addresses three major areas. First, it highlights how signature verification is taken differently in the above mentioned communities and why this gap is increasing. Various factors that widen this gap are discussed with reference to some of the recent signature verification studies and probable solutions are suggested. Second, it discusses the state-of-the-art evaluation and its problems as seen by FHEs. The real evaluation issues faced by FHEs, when trying to incorporate automatic signature verification systems in their routine casework, are presented. Third, it reports a standardized evaluation scheme capable of fulfilling the requirements of both PR researchers and FHEs.

***Keywords*-Signatures verification, terminology, standardization, evaluation, forensic casework, handwriting examiners, signature datasets**

I. Introduction

The research community is interested in signature verification for centuries. The technological revolution that came by the emergence of computers has shifted this interest towards automatic verification of signatures. Today, industry in general and FHEs in particular look forward for automatic signature verification. It will assist FHEs in performing their routine signature verification tasks efficiently and effectively. While there is an increasing demand of automatic systems in forensic handwriting examination departments, there are certain barriers/gaps between the two communities, i.e., PR community and FHEs community, starting from the terminology till the evaluation of outcomes. This paper will highlight a few of the most important of these gaps and will suggest probable solutions.

Note that forensic scenarios have some inbuilt complexities like signature samples taken from torn or muddy papers, clothes, etc., which are not considered here. We will focus on daily routine cases of FHEs, where signature samples are available similarly to PR research and yet FHEs find application of automatic signature verification systems nearly impossible. The specific problems addressed are non-accessible data sets and non-representative data, difference in terminologies used by different PR researchers, the output reported by automatic systems and the current evaluation scheme.

This paper can be viewed as an effort to standardize the evaluation criteria for signature verification. The standardized criteria will be useful for both the PR researchers and FHEs. It is a similar effort as done in the field of image binarization by [1] where various binarization evaluation metrics were discussed.

The rest of this paper is organized as follows. In Section II signature verification is defined with respect to PR researchers. Section III defines signature verification with respect to FHEs. This is required to highlight the basic differences and similarities among the definitions of various modalities in the two communities. Section IV makes the core of this paper where we move towards standardization. There the most important barriers to the application of automatic systems in forensic departments are presented. Then each of these barriers is discussed in detail and it's probable solutions are suggested. A special attention is given to the third and the fourth barriers, i.e., state-of-the-art output and state-of-the-art evaluation, respectively. Section V summarizes the paper and suggests some future work.

II. Automatic Signature Verification: PR-view

Today the PR community moves by defining the automatic signature verification as a two-class pattern classification problem [2]. Note that in earlier PR studies it was defined differently where PR researchers also considered other genres of signatures such as, disguised signatures [3], see Section III. As a two class classifier, an automated system has to decide whether or not a given signature belongs to a referenced authentic author. If a system could find enough evidence of genuine authorship from the questioned signature's feature vector, it considers the signature as genuine; otherwise it declares the signature as forged. Note that in the PR community various types of forgeries are studied and in fact sometimes the same forgery type is termed differently by different researchers.

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.205

We are only reporting the most common forgery types as studied by PR researchers. These are,

1) Random Forgery: genuine signature of any writer other than the authentic author.
2) Simple/Casual Forgery: the forger only knows the name of the authentic author.
3) Simulated Forgery: produced by inexperienced forger after practicing unrestricted number of times.
4) Skilled Forgery: produced by experienced forger, usually a calligrapher, after practicing unrestricted number of times.

Moreover, automated signature verification is divided into, online and offline, depending on the mode of the handwritten input. If both spatial and temporal information are available to a system, verification is performed on online data. In the case where temporal information is not available and a system can only utilize the spatial information gleaned through scanned or camera captured documents, verification is performed on offline data [2].

III. Signature Verification: FHEs-view

FHEs do not view signature verification as a two class classification problem [4]. It involves various other genres of natural and unnatural handwriting. Due to space limitations, we will focus only on the main types of non-genuine signatures relevant to discussion at hand.

1) Disguised Signatures: *not* a forgery, but an authentic author imitates his/her own signature to make it look like a forgery so that it can be denied at a later time. Note that disguised signatures are an unnatural signing behavior, yet they are from the authentic author. So, if we are to establish authorship, disguised signatures must lay in positive authorship category.
2) Simple Forgery: the forger knows the original signatures (seen for sometime) and forges without practice. This is in contrast to the PR definition of simple forgery where a forger may only know the name of an authentic author.
3) Skilled Forgery: the forger knows the original signatures and forges after practicing unrestricted number of times.

Note that FHEs take the terms simulated and forged in the same meaning [5]. The forgeries/simulations can either be made free hand or traced. Whether traced or free hand, they can be simple or skilled. Furthermore, note that the term "random forgery" does not appear here. This is because a random forgery, as defined by PR researchers, is considered as a fictitious case by FHEs and hence they do not study it [5].

IV. Moving Towards Standardization: Bridging the Gaps

In the following section we highlight various gaps/barriers which must be considered for development of a common understanding between the PR researchers and FHEs. These include,

- Non-accessible datasets and non-representative data.
- Different terminology and modalities.
- State-of-the-art output of automatic systems.
- State-of-the-art evaluation.

A. Non-accessible datasets and non-representative data

Many PR systems are not trained/tested on publicly available data and therefore the experiments are not repeatable/verifiable. Due to this the FHEs can never be sure of which systems can potentially be better applied to their casework. Furthermore, a majority of the state-of-the-art signature verification systems are built, tested, and optimized for data that are not a representative for data faced in forensic cases. These PR data usually contain various fictitious signatures, such as "random forgeries" as discussed above.

Probable Solution

To bridge this gap the PR researchers should use data that are publicly available preferably collected by FHEs in forensic like situations. Today a large amount of such data are publicly available, such as the data from various signature verification competitions jointly organized by PR researchers and FHEs. These include SigComp2009 [6], 4NSigComp2010 [4], SigComp2011 [7][1]. Having different automated systems that report results on the same data sets may provide a comparative analysis of their performances.

If application specific data are collected for special purposes they should be unbiased and have statistical significance. Moreover the following information is required.

- Data collection procedure.
- Any specific restrictions applied while collection, any errors occurred and corrective measures taken.
- When and if they will be publicly available.

B. Different Terminology and Modalities

PR researchers and FHEs define some of the signature verification modalities differently, e.g., the term *Simple Forgery* (refer to the Sections II and III). Moreover, different PR researchers sometimes give the same name to somewhat different signature modalities. Some examples of such mismatch include [8], [9] and [14]. In addition to that, in some cases a lot of PR research reveals results that are trivial/irrelevant with respect to forensic casework, e.g., a common practice of PR researchers is to report random forgeries but they are fictitious in view of forensic experts. If random forgeries are included in a test set while evaluating the results of a system, a system having very low error rate may still not be suitable for forensic casework. On the other hand a system having a high error rate but considering skilled forgeries may yield

[1] available at http://www.iapr-tc11.org/mediawiki/index.php/Datasets_List

Table I
RESULTS OF SOME RECENT SIGNATURE VERIFICATION SYSTEMS. HERE, F=FORGERY, G=GENUINE SIGNATURE, RF=RANDOM FORGERY (NOT RELEVANT FOR FORENSIC CASEWORK), SF=SIMPLE FORGERY, SK=SKILLED FORGERY, SM=SIMULATED FORGERY, AND T=TOTAL NUMBER OF SIGNATURES.

Study	Database			FAR(%)	FRR(%)
[8]	320(G)	320(F)	640(T)	0.11	0.02
[9]	300(G)	300(F)	600(T)	4.16	7.51
[10]	980(G)	980(F)	1960(T)	0.01(RF),4.29(SF),19.80(SK)	2.04
[11]	300(G)	600(F)	900(T)	4.41(RF),1.67(SF),15.67(SM)	10.33
[12]	2400(G)	0(F)	2400(T)	0.64	1,17
[13]	500(G)	0(F)	500(T)	9.81	3

better results in forensic casework. Examples are reported in Table I. Note that these examples are from PR literature and are presented here just to highlight the difficulty the FHEs face when viewing these results where different types of signature modalities are either differently defined or in some cases are combined with each other while reporting the overall system performance.

As given in Table I, Systems [10] and [11] are producing higher error rates with skilled and simulated forgeries than other systems which either do not specify the types of forgeries considered, e.g., Systems [8], [9], or do not use forgeries in evaluation like [12], [13]. By seeing these results an FHE cannot say anything with certainty about which system will perform better in real forensic casework. However Systems [10] and [11] are reporting their results on skilled forgeries separately which is worth more for an FHE.

Probable Solution

Settling down on common definition would favor the application of automatic systems in real forensic casework [5]. It is suggested here that the two communities may use the following terms,

- Genuine: for authentic signatures.
- Forged: for simulated signatures.
- Simple forgery: a forgery where actual signatures are known but forgery is produced without any practice.
- Skilled forgery: same as simple forgery but produced after practice.

If random forgeries are used for some specific purposes, they must be separated from the other types and should not affect the overall evaluation. Moreover, other types of signing behaviors as studied by FHEs, e.g., disguised signatures, should be focused by PR researchers. The datasets of 4NSigComp2010 and 4NSigComp2012[2] can be used for this purpose.

C. State-of-the-Art Output of Automatic Systems

What should an automated signature verification system output in order to be successfully applicable in forensic casework? This is a substantial question for both PR researchers and FHEs.

[2]available on the TC-11 page (see footnote 1) from May 2012

The output produced usually by automated systems is not acceptable for presentation in the courts thereby making the use of automatic systems nearly impossible for FHEs [7]. Traditionally, automated signature verification systems report their decisions in a boolean manner, i.e., if enough evidence of a forgery is present, a system reports a reject, otherwise an accept. Though this is quite objective and may be significant in some fields like real time application, e.g., banking, but a boolean answer of *genuine* or *forged* is not adequate for the FHEs. They are interested to exactly know how close is a questioned signature to a genuine signature when it is declared as forged and vice versa.

To bridge this gap, automated systems usually provide some sort of similarity score between 0 and 1, e.g., probability values. Here a value near 0 represents a forgery and a value near 1 represents genuine authorship. This again is inadequate for forensic casework. This is because mere scores/probability values in themselves raise many questions for FHEs and courts. How are these values related to the authorship (genuine or forged) and among themselves; How to compare different systems producing different values for the same questioned signature; How would an FHE establish that a value of 0.2 produced by one automated system is still more close of being genuine signature than a value of 0.4 produced by another system; How would these sort of outputs be defended in courts?

Moreover, FHEs are interested to know the features contributing to the output. They would like to consider the features' uniqueness/rarity in a population, e.g., how rare is the style of writing a special character in a population? This information impacts the overall evaluation of an FHE while examining a signature sample. But how would that relate to an automated system?

Probable Solution

A probable solution is that the automated systems should produce some continuous similarity/difference score s (that may vary between any two extremes like, 0.001 to 1000)which would be converted into evidential value/Likelihood Ratios (LR) according to the Bayesian approach [15].

The idea of this solution is given in Figure 1. Here the score s is computed by comparison between the questioned

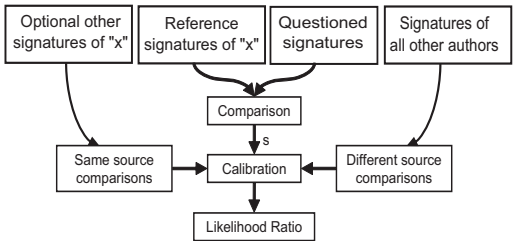

Figure 1. Evaluation Scheme.

signature and the reference signatures. In addition to that the different source comparison is performed by considering the signatures of all other authors available in the training set. The same source comparison is also performed if other signatures of the referenced author are available. The scores can be converted into LRs by a so-called calibration procedure [16], such as the one implemented in the FoCal toolkit[3]. The LR is actually the ratio of the probabilities of finding that score s when questioned signature is genuine and when it is forged. Such LR values will suffice the needs of FHEs especially by considering the features' rarity in a population.

Note that so-far the computation of some kind of likelihood ratios has been realized in different ways only in some tools and frameworks, such as CEDAR-FOX [17][4] and WANDA project [19], but in general the PR community has not adopted the likelihood ratios at large. This makes the application of the majority of state-of-the-art PR methods impossible in forensic casework.

Furthermore, PR researchers are usually not interested in continuous values like likelihood ratios. They usually demand an objective indicators of a system's performance. In the next section we will first explain the state-of-the-art of PR evaluation. Then suggest how the above mentioned likelihood ratios may be converted to depict an objective measure to suit the needs of PR researchers at the same time.

D. State-of-the-Art Evaluation

The results of signature verification systems are evaluated differently by PR researchers and FHEs. In PR the following terms are widely used to report on the evaluation of automatic signature verification systems.

- Accuracy: measures the percentage of correctly classified signatures with respect to all signatures under investigation [20]. When only accuracy is used, a system that votes by chance may show higher accuracy if, for example, there are unequal number of genuine and forged signatures in a dataset. To rectify this, often FRR and FAR are considered.
- False Rejection Rate (FRR): Also known as *miss probability* or *Type I error*. It is the rate at which genuine signatures are classified as forged by a system.
- False Acceptance Rate (FAR): Also known as *false alarm probability* or *Type II error*. It is the rate at which forged signatures are classified as genuine by a system. Both FRR and FAR are usually given in percentage.
- Receiver Operating Characteristic (ROC) curve: FRR and FAR can be computed at any given threshold but in order to view the complete behavior of a signature verification system an ROC-curve is often plotted. This curve represents FRR and FAR at all possible thresholds for a system[5].
- Area Under Curve (AUC): It represents the probability that a system gives higher value to a randomly chosen genuine signature as compared to a randomly chosen forgery. As the name suggests, the smaller the area under ROC-curve, the better is a system's performance.
- Equal Error Rate (EER): The point on the ROC-curve where FRR equals FAR.
- Average Error Rate (AER): It is the mean of FRR and FAR. Usually used when no decision threshold can be adjusted, e.g., on a final test set when trying to assess the performance without adjusting any parameter.
- Detection Error Trade-off (DET) curve: It is a variant of the ROC curve that is plotted by taking FRR and FAR on a logarithmic scale.

The current state-of-the-art of evaluation, as given above, is not adequate for FHEs. In many circumstances an FHE reports a continuous measure of evidential value to the court. In fact an FHE is included in a judicial investigation to facilitate the court by analyzing handwritten text/signatures and reporting the evidence of a forgery or genuine authorship. The court does not demand a decision from an FHE, rather a continuous measure of similarity or difference. This is because a pure classification, as done in PR, cannot be combined with other circumstantial evidence, e.g., opportunity, motive, fingerprints, etc. in a legally acceptable way.

Since the current evaluation methods primarily evaluate the systems' performances on the basis of correct or wrong classification they are inadequate to fulfill the needs of FHEs. Therefore we need an alternative evaluation scheme that would serve the following two purposes.

- First, fulfill the demands of FHEs and enable them present the results of automated systems in courts.
- Second, objective enough so that to enable PR researchers compare the performance of their systems in an objective/direct manner preferably also representable as a single score.

[3]http://focaltoolkit.googlepages.com

[4]the practicability of CEDAR-FOX is discussed in [18]

[5]as an alternative, the FAR could also be plotted against True Accept Rate (TAR)

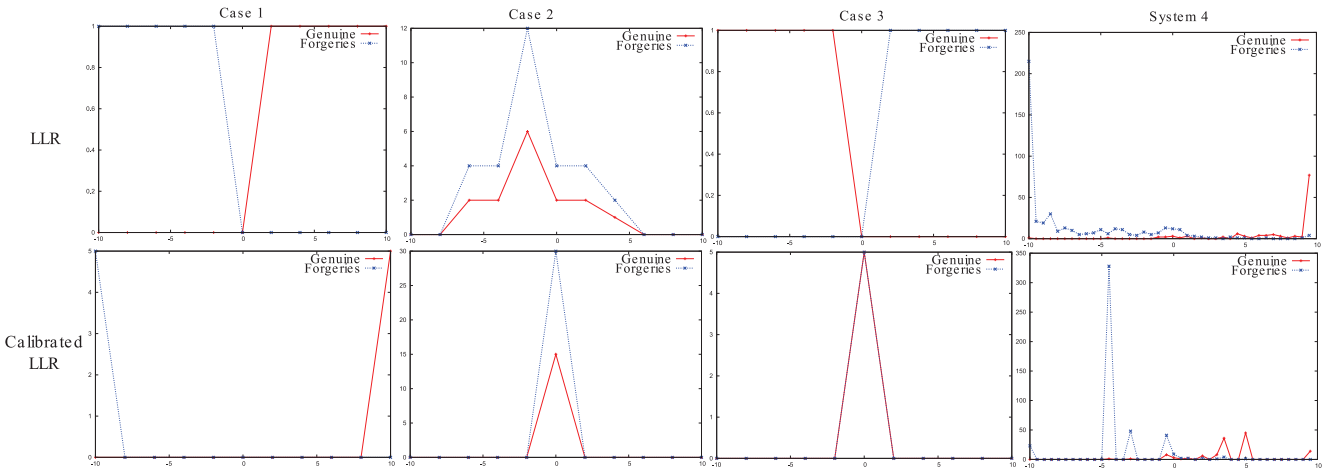

Figure 2. LLR curves before (on top) and after (on bottom) calibration, possible extreme cases (Case 1 to Case 3), and results of System 4 of Table II. X-axis: (Optimized) Log Likelihood, Y-axis: No. of Occurences.

Probable Solution: Standardized Performance Evaluation Scheme

As suggested previously, automatic signature verification systems should output continuous scores that can be converted into LRs or Log LRs (LLRs) by various calibration procedures.

These conversions need to be monotonic thereby not affecting the discrimination between the genuine and forged signatures as suggested by an automated system. After these conversions, the signatures having weak evidence of forgery generally do not lead to high absolute values of LRs, while the signatures with a strong evidence of forgery lead to high absolute values of LRs.

We have already tested this evaluation scheme and applied it successfully in the SigComp2011 signature verification competition. Table II shows some of the results received on the Chinese online dataset[6]. Note that this evaluation scheme suggests the PR researchers to report results in the form of LLRs and their corresponding cost $\widehat{C}_{llr}$. The LLR would help FHEs in presenting the results of automated systems in courts. Furthermore, $\widehat{C}_{llr}^{min}$ is calculated that is the minimum possible value of $\widehat{C}_{llr}$ as suggested in [16]. This $\widehat{C}_{llr}^{min}$ value can be used as a final assessment score of a system's performance in PR research.

As depicted in Table II, the system with the best FRR and FAR also has the best value of $\widehat{C}_{llr}^{min}$, i.e., (minimum value). But from here we cannot generalize that a system having better FRR and FAR will always have better $\widehat{C}_{llr}^{min}$. For example, System 9(b) performs quite well at the FRR/FAR scale but has the worst $\widehat{C}_{llr}^{min}$. This might be explained by the fact that even a few misleading answers with high score can spoil the overall performance of $\widehat{C}_{llr}^{min}$. Note that this reflects the practice, i.e., a system should not produce a high likelihood for a wrong decision as this might result in wrong judgment with severe outcomes.

[6]all datasets are publicly available at TC-11 page

Table II
RESULTS ON CHINESE ONLINE DATASET OF SIGCOMP2011

System	Accuracy(%)	FRR	FAR	$\widehat{C}_{llr}$	$\widehat{C}_{llr}^{min}$
1(b)	84.81	12.00	16.05	0.56	0.35
4	93.17	6.40	6.94	0.41	0.22
6(b)	82.94	16.80	17.14	1.05	0.50
7(b)	85.32	13.60	14.97	0.90	0.46
9(b)	80.89	9.26	8.14	6.21	0.73

To further clarify this evaluation scheme three extreme cases that can occur while following this scheme in evaluation are presented in the Figure 2. The distributions of the LLRs of genuine signatures (target values) and forgeries (non-target values) are depicted in red lines and blue dotted lines, respectively. These distributions can be interpreted as follows. The curves on the top represent the non calibrated evidential values, while the respective curves on the bottom are produced after calibrating with the Focal tool kit. The farther the tar-curve goes to the left, the higher would be the cost of this misleading decision. Similarly, the farther the nontar-curve goes to the right, the higher would be the cost. For optimal performance and thereby for minimum value of $\widehat{C}_{llr}^{min}$, the two curves must be optimally separated.

In Case 1 the curves are perfectly separated and lay on their desired sides. Therefore, the calibrated LLRs have the minimum cost, thus the $\widehat{C}_{llr}^{min}$ will be equal to 0. Cases 2 and 3 present a perfectly non-distinguishing and an always misclassifying system, respectively. In both of these cases, the $\widehat{C}_{llr}^{min}$ will be equal to 1 that is the maximum cost. As a reference, the LLR curves for System 4 of Table II are given in Figure 2 on the right side. Note that as there was a wrong decision with a high LLR (marked with a green circle), the calibrated LLRs have lower absolute values.

V. Summary and Future Directions

In this paper we have presented various barriers/issues that hinder the application of automatic signature verification systems in real forensic casework and proposed solutions. To do so we shortly summarized the view points of the two communities in general about signature verification. Then we described barriers between the two communities in detail and suggested solutions.

In future it is hoped that both the PR and FHE communities would move together to further enhance the scope of their collaborative work. PR experts are encouraged to develop system also targeting the needs of FHEs. In the meanwhile FHEs are encouraged to use more and more automated systems in their every day casework and provide feedback to PR researchers. This is substantial since improvements in the current systems are only possible if they are exposed to tackle rigorous real world signature verification scenarios. Furthermore, the FHEs should make more forensically relevant data publicly available. This will be a first step towards a common goal by the two communities, i.e., the application of automated systems to assist in solving real forensic handwriting analysis cases.

References

[1] E. H. B. Smith, "An analysis of binarization ground truthing," in *9th IAPR Int. Workshop on Document Analysis Systems*, ser. DAS '10. New York, USA: ACM, 2010, pp. 27–34.

[2] D. Impedovo and G. Pirlo, "Automatic signature verification: The state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 5, pp. 609–635, Sep. 2008.

[3] R. Plamondon and G. Lorette, "Automatic signature verification and writer identification – the state of the art," *Pattern Recognition*, vol. 22, pp. 107–131, 1989.

[4] M. Liwicki, C. E. van den Heuvel, B. Found, and M. I. Malik, "Forensic signature verification competition 4NSigComp2010 - detection of simulated and disguised signatures," in *12th Int. Conf. on Frontiers in Handwriting Recognition. (ICFHR-2010), November 16-18, India*, 2010, pp. 715–720.

[5] M. Liwicki, M. Blumenstein, B. Found, C. E. van den Heuvel, C. Berger, and R. Stoel, Eds., *Proceedings of the 1st Int. Workshop on Automated Forensic Handwriting Analysis*, vol. 768. CEUR-WS, 2011. [Online]. Available: CEUR-WS.org/Vol-768/

[6] V. L. Blankers, C. E. van den Heuvel, K. Y. Franke, and L. G. Vuurpijl, "ICDAR 2009 signature verification competition," in *Int. Conf. on Document Analysis and Recognition*, 2009, pp. 1403–1407.

[7] M. Liwicki, M. I. Malik, C. E. van den Heuvel, X. Chen, C. Berger, R. Stoel, M. Blumenstein, and B. Found, "Signature verification competition for online and offline skilled forgeries SigComp2011," in *11th Int. Conf. on Document Analysis and Recognition*, 2011, pp. 1480–1484.

[8] E. Ozgunduz, T. Senturk, and M. E. Karsligil, "Off-line signature verification and recognition by support vector machine," in *13th European Signal Processing Conf. (EUSIPCO 2005)*, Sep. 2005.

[9] S. Kumar, K. B. Raja, R. K. Chhotaray, and S. Pattanaik, "Off-line signature verification based on fusion of grid and global features using neural networks," *Int. Journal of Engineering Science and Technology*, vol. 2, pp. 7035–7044, 2010.

[10] L. Cordella, P. Foggia, C. Sansone, F. Tortorella, and M. Vento, "A cascaded multiple expert system for verification," in *Multiple Classifier Systems*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2000, vol. 1857, pp. 330–339.

[11] C. Santos, E. J. R. Justino, F. Bortolozzi, and R. Sabourin, "An off-line signature verification method based on the questioned document expert's approach and a neural network classifier," in *9th Int. Workshop on Frontiers in Handwriting Recognition, IWFHR-9 2004.*, 2004, pp. 498–502.

[12] A. El-yacoubi, E. J. R. Justino, R. Sabourin, and F. Bortolozzi, "Off-line signature verification using hmms and cross-validation," in *Proc. of the IEEE Workshop on Neural Networks for Signal Processing*, 2000, pp. 859–868.

[13] H. Baltzakis and N. Papamarkos, "A new signature verification technique based on a two-stage neural network classifier," *Engineering Applications of Artificial Intelligence*, vol. 14, no. 1, pp. 95–103, Feb. 2001.

[14] E. Justino, E. J. R. Justino, F. Bortolozzi, and R. Sabourin, "Off-line signature verification using hmm for random, simple and skilled forgeries," in *Simple and Skilled Forgeries, ICDAR 2001, Int. Conf. on Document Analysis and Recognition*, 2001, pp. 1031–1034.

[15] J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, D. Ramos-Castro, and J. Ortega-Garcia, "Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems," *Forensic Science Int.*, vol. 155, no. 2-3, pp. 126–140, 2005.

[16] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[17] S. N. Srihari, B. Zhang, C. Tomai, S. Lee, Z. Shi, and Y. C. Shin, "A system for handwriting matching and recognition," in *Symposium on Document Image Understanding Technology*, 2003, pp. 67–75.

[18] R. J. VERDUIJN, C. E. van den Heuvel, and R. D. STOEL, "Forensic requirements for automated handwriting analysis systems," in *The 15th Int. Graphonomics Society Conf. (IGS2011)*, Live Aqua Cancun, Mexico, June 2011, pp. 132–135.

[19] K. Franke, L. Schomaker, C. Veenhuis, L. Vuurpijl, and I. Erp, M. van Guyon, "WANDA: A common ground for forensic handwriting examination and writer identification," *ENFHEX News*, pp. 23–47, 2004.

[20] T. Fawcett, "An introduction to ROC analysis," *Pattern Recogn. Lett.*, vol. 27, pp. 861–874, June 2006.